Bioinformatics, 37(20), 2021, 3491–3493
doi: 10.1093/bioinformatics/btab373
Advance Access Publication Date: 12 May 2021
Original Paper

OXFORD

Structural bioinformatics

ProteoMill: efficient network-based functional analysis portal for proteomics data

Martin Rydén [1,*], **Martin Englund**[1,†] **and Neserin Ali**[2,†]

[1]Department of Clinical Sciences Lund, Orthopedics, Clinical Epidemiology Unit, Lund University, SE-22185 Lund, Sweden and [2]Department of Clinical Sciences Lund, Rheumatology and Molecular Skeletal Biology, Lund University, SE-22184 Lund, Sweden

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the last two authors should be regarded as Joint Last Authors.

Associate Editor: Pier Luigi Martelli

Received on February 7, 2021; revised on April 25, 2021; editorial decision on May 6, 2021; accepted on May 11, 2021

Abstract

Summary: Functional analysis has become a common approach to incorporate biological knowledge into the analysis of omics data, and to explore molecular events that govern a disease state. It is though only one step in a wider analytical pipeline that typically requires use of multiple individual analysis software. There is currently a need for a well-integrated omics analysis tool that performs all the steps. The ProteoMill portal is developed as an R Shiny application and integrates all necessary steps from data-upload, converting identifiers, to quality control, differential expression and network-based functional analysis into a single fast, interactive easy to use workflow. Further, it maintains annotation data sources up to date, overcoming a common problem with use of outdated information and seamlessly integrates multiple R-packages for an improved user-experience. The functionality provided in this software can benefit researchers by facilitating the exploratory analysis of proteomics data.

Availability and implementation: ProteoMill is available at https://proteomill.com.

Contact: martin.ryden@med.lu.se

1 Introduction

The large amounts of data generated from omics experiments have stressed the need for methods to reveal and extract critical components of dynamic biological systems in a readable manner, which connects to the specific study question. Expression data that are derived from high throughput analysis have multiple levels of biological features connected to it. In a real biological environment, the physical, genetic, regulatory and functional properties of a molecular set work together in a response to environmental stimuli. Holistically evaluating these attributes is a way to reveal the intercommunication between these properties and to provide a biological context. However, this task encompasses some impending challenges, including differences in biomolecule identification, data dimensionality reduction, biological contextualization, statistical analysis and data visualization and this differs among the various types of individual datasets.

Existing omics analysis tools are typically specialized for individual parts of the analysis workflow and differences in data format standards means the tools do not integrate well when used as part of an analysis workflow. This requires the researcher not only to have knowledge of the different individual software, but also knowing how to format the generated output from one software for use in the next software. This often poses a time-consuming task, particularly for researchers with little computational experience or little experience with the software(s) in question and is prone to errors.

Omics analysis platforms such as Perseus (Tyanova and Cox, 2018) and Qlucore (Qlucore, 2021) offer thorough analytical and explorative features, but require users to download and install their software and is not open source. While there are many existing web-based omics tools which are able to perform individual parts of an analysis workflow (Efstathiou *et al.*, 2017; Kuleshov *et al.*, 2016; Luo *et al.*, 2017; Merico *et al.*, 2010; Perlasca *et al.*, 2019; Schweppe *et al.*, 2017; Zheng and Wang, 2008), many lack the ability to perform complete pipelines in fast, interactive web-environments. Reimand *et al.* lists the protocols and time consumption for popular enrichment software, with the time expense ranging from minutes to several hours (Reimand *et al.*, 2019). In contrast, the run time for ProteoMill functions are a few seconds at the most, as described in Table 1.

Another important but often overlooked aspect for generating reliable and biologically relevant results is the quality of annotation data, and, by extension, a tool's ability to maintain annotation data sources up to date. Lina Wadi *et al.* reported that 67% of publications in their survey referenced software using outdated annotation data (Wadi *et al.*, 2016). Web-based tools have an inherit advantage in that back-end data sources can be dynamically updated without requiring manual action by the user (such as downloading and installing software).

Analysis of proteomic data faces additional challenges (Kirik *et al.*, 2012). Different gene- and protein level identifier types are utilized in the various omics tools, which often require the researcher to convert between identifier types before proceeding to the next step of the analysis. This can result in loss of data since there can exist one-to-many mappings between two identifier types or that an identifier cannot be mapped between two identifier types (Reimand *et al.*, 2019). Furthermore, a frequent concern in mass spectrometry-derived data is the abundance of missing values (Lazar *et al.*, 2016; Wang *et al.*, 2017).

Thus, a tool that could help to transform the biological research into integrated framework is preferred. The aim of this study is to describe a newly developed software that addresses many of the existing shortcomings. The fundamental concepts of this software are to provide sets of well-integrated, easy-to-use and to a large extent automated functions for exploratory analysis of proteomic data.

2 Materials and methods

2.1 Architecture

ProteoMill runs as a web application using Shiny Server and is hosted on Amazon Web Services. The software is developed in R (version 3.6.1) and the interface was created using the R-package Shiny (Chang *et al.*, 2021) and shinydashboard (Chang and Ribeiro, 2018) (version 0.7.1) with a customized CSS theme. Animations were created using jQuery and the library animejs. Plotly (Plotly Technologies Inc., 2015), ggplot2 (Wickham, 2009), heatmaply (Galili *et al.*, 2018), networkD3 (Allaire *et al.*, 2017) and visNetwork (Almende et al., 2019) were used for plotting.

2.2 Identifier conversion

The Bioconductor packages AnnotationDbi (Pagès *et al.*, 2020) and ensembldb (Rainer *et al.*, 2019) was used for converting between identifiers. The identifier type of the user's uploaded data is automatically recognized and converted to four different identifier types (where applicable). This way, the user can choose to display protein labels as any of the five identifier types, but do not need to worry about manually converting between identifiers.

2.3 Data quality control

Principal component analysis (PCA) was implemented using the R-package stats. Another package, mixOmics, was used for multilevel PCA.

2.4 Differential expression analysis

Two R-packages, limma (Ritchie *et al.*, 2015) and DESeq2 (Love *et al.*, 2014) were implemented for differential expression analysis. Each package is commonly used for fitting gene-wise linear models to expression data. limma was originally developed with a primary focus on the analysis of microarray data, while DESeq2 for the analysis of RNA-seq data and is based on the negative binomial distribution.

Differential expression analysis is conducted by specifying two contrasts and choosing a paired or non-paired design. The results are evaluated by inspecting the table in the 'Differential expression' tab.

The results are displayed as estimated by the specific software, using the software's default settings for shrinkage parameters, correction for multiple testing, significance level and etc. For example, the correction for multiple testing is done using the Benjamini–Hochberg method and is applied to the tests performed within one run of the analysis and not with respect to all tests performed within one family of hypotheses in a study, which sometimes may be misleading (Ranstam, 2016). The user needs to verify if these setting are appropriate for the specific analysis done.

2.5 Functional enrichment and network analysis

The hypergeometric distribution was used to calculate the probability of protein list overlap.

$$P = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} \quad (1)$$

In this formula, N is the total number of proteins in the background distribution, M is the number of proteins in the background distribution annotated to a pathway, n is the total number of selected proteins of interest and x are the proteins of interest annotated to a pathway.

Pathway data and interaction data are dynamically collected from Reactome (Fabregat *et al.*, 2018) and STRING (Szklarczyk *et al.*, 2015) (https://reactome.org/download-data). MD5sum hashes are used to ensure that the local database is up to date.

For each entry in the main pathway data file, the top-level parent pathway was annotated. This was done by creating a directed acyclic graph object using the R-package igraph (Csárdi and Nepusz, 2006).

2.6 Data sources

An important aspect of this software is to maintain data sources up to date. This is done by using an automated workflow at a bi-monthly interval. Data are collected from the two primary data sources, Reactome (Fabregat *et al.*, 2018) for pathway data and STRING (Szklarczyk *et al.*, 2015) for protein interaction data. These data are then structured to a predefined format, making it possible to integrate them in the analysis.

3 Results

The presented software, ProteoMill, proposes a unique approach to conducting explorative analysis of proteomics data. The data visualization capabilities present in this software are designed to make it possible even for researchers without any particular computational training to gain insights about the biological meaning of their data. Many of the graphical components are interactive, which is a useful feature for analysing protein interactions and selecting subnetworks of interest.

A common goal in many of ProteoMill's functionalities is to reduce data complexity, and to provide a framework for extracting elements of biological relevance. PCA reduces a dataset of hundreds or thousands of expression datapoints into a single datapoint for each condition, plotted in 2–3 principal components, which in turn describes the dimensions with largest variability. The datapoints cluster together based on the similarity of their expression profiles.

Categorizing proteins into biological entities, described as pathways, is another way to reduce complexity and make sense of one's data. Network graphs produced from interaction data can be difficult to interpret. In ProteoMill, pathways are used to categorize and label groups of interacting proteins, and as a way to inspect subnetworks based on these common biological themes.

The integrated enrichment- and network analysis provides a way for users to simultaneously explore functional analysis output and interaction data, and this feature has been specifically designed to easily identify and select subnetworks of interest for further analysis.

3.1 Reproducibility

ProteoMill supports the use of reproducibility tokens as a simple way to load settings and database versions from a prior session. The token contains information about all user defined settings that affect the outcome of the analysis—every statistical result and its graphical representations. The token also contains an MD5sum hash for the uploaded dataset and warns the user if the uploaded file is not identical to the file used in the previous session.

3.2 Performance

To assess the performance of ProteoMill, we measured the execution speed of its most prominent functions directly on the server (Table 1), using a publicly available dataset consisting of 12 samples and 12 320 proteins (Wertheim *et al.*, 2009). The time elapsed for rendering plots depends on the client-side machine and browser. The column labeled

Table 1. Benchmark results of ProteoMill functions

Function	Exec. time (ms)	Total time (ms)
Upload data	2560	—
Set missing value cutoff	90	—
Differential expression (limma)	215	—
Differential expression (DESeq2)	5042	—
Pathway over-representation	367	—
Interaction network[a]	53	1041
Interaction network[b]	58	5580

[a]Using 162 nodes and 582 edges.

[b]Using 1356 nodes and 17 718 edges.

'Exec. Time' describe the elapsed time of server-side calculations/data sub-setting operations and the column 'Total time' also describes the rendering time as measured on a 2018 MacBook Pro (2.2 GHz 6-Core Intel Core i7).

4 Discussion

The integrated features in this software provide powerful visualization strategies for the exploration of omics data, with a particular focus on the management and manipulation of proteomics data. By using this platform, researchers can expect to discover biologically relevant rendering of their data through results aggregated from reliable and up-to-date data sources.

The software offers innovative strategies to interactively explore quantitative proteomics data in a comprehensive workflow from data-upload to network analysis. It has a strong focus on well-maintained data sources, computational efficiency and user-friendliness.

Importantly, ProteoMill utilizes many existing R packages for statistical analysis and pathway annotation that are standard in the field. However, these methods are strongly focused on estimation of P-values and classifications of results based on P-value thresholds. This is an unfavorable approach to use of statistical methods and there is a need to move further in better estimation methods and expressing uncertainty (Benjamin *et al.*, 2018).

Acknowledgements

The authors would like to thank Aleksandra Turkiewicz, PhD for helpful discussions and critical reading of this paper.

Funding

This work was supported by 'The Swedish Research Council (Vetenskapsrådet), grant number 542-2014-2348' and 'Kockska stiftelsen för medicinsk forskning (Fromma)'.

Conflict of Interest: none declared.

Data availability

The tool itself is available on https://proteomill.com/. All demonstrational data can be downloaded directly from the website by clicking Demo data, selecting a dataset and clicking 'Download'.

Code availability

The source code is available at https://github.com/martinry/ProteoMill/.

References

Allaire,J.J. *et al.* (2017) NetworkD3: D3 JavaScript network graphs from R. Last accessed 23 April, 2021.

Almende, B.V. *et al.* (2019) visNetwork: network visualization using 'vis.js' Library. Last accessed 23 April, 2021.

Benjamin,D.J. *et al.* (2018) Redefine statistical significance. *Nat. Hum. Behav.*, **2**, 6–10.

Chang,W. *et al.* (2021) Shiny: web application framework for R. Last accessed 23 April, 2021.

Chang,W. and Ribeiro,B.B. (2018) shinydashboard: create dashboards with 'Shiny'. Last accessed 23 April, 2021.

Csárdi,G. and Nepusz,T. (2006) The igraph software package for complex network research.

Efstathiou,G. *et al.* (2017) ProteoSign: an end-user online differential proteomics statistical analysis platform. *Nucleic Acids Res.*, **45**, gkx444.

Fabregat,A. *et al.* (2018) The Reactome pathway knowledgebase. *Nucleic Acids Res.*, **46**, D649–D655.

Galili,T. *et al.* (2018) Heatmaply: an R package for creating interactive cluster heatmaps for online publishing. *Bioinformatics*, **34**, 1600–1602.

Kirik,U. *et al.* (2012) Multimodel pathway enrichment methods for functional evaluation of expression regulation. *J. Proteome Res.*, **11**, 2955–2967.

Kuleshov,M.V. *et al.* (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.*, **44**, W90–W97.

Lazar,C. *et al.* (2016) Accounting for the multiple natures of missing values in label-free quantitative proteomics data sets to compare imputation strategies. *J. Proteome Res.*, **15**, 1116–1125.

Love,M.I. *et al.* (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.

Luo,W. *et al.* (2017) Pathview Web: user friendly pathway visualization and data integration. *Nucleic Acids Res.*, **45**, W501–W508.

Merico,D. *et al.* (2010) Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS One*, **5**, e13984.

Pagès,H. *et al.* (2020) AnnotationDbi: manipulation of SQLite-based annotations in Bioconductor. Last accessed 23 April, 2021.

Perlasca,P. *et al.* (2019) UNIPred-Web: a web tool for the integration and visualization of biomolecular networks for protein function prediction. *BMC Bioinformatics*, **20**, 422.

Plotly Technologies Inc. (2015) *Collaborative Data Science*. Montréal, QC: Plotly Technologies Inc.

Qlucore AB. (2021) Qlucore omics explorer.

Rainer,J. *et al.* (2019) ensembldb: an R package to create and use Ensembl-based annotation resources. *Bioinformatics*, **35**, 3151–3153.

Ranstam,J. (2016) Multiple P-values and Bonferroni correction. *Osteoarthr. Cartilage*, **24**, 763–764.

Reimand,J. *et al.* (2019) Pathway enrichment analysis and visualization of omics data using g: profiler, GSEA, Cytoscape and EnrichmentMap. *Nat. Protoc.*, **14**, 482–517.

Ritchie,M.E. *et al.* (2015) Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47.

Schweppe,D.K. *et al.* (2017) BioPlex display: an interactive suite for large--scale, AP-MS protein-protein interaction data. *J Proteome Res.* **17**, 722–726.

Szklarczyk,D. *et al.* (2015) STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.*, **43**, D447–D452.

Tyanova,S. and Cox,J. (2018) Cancer systems biology: methods and protocols. *Methods Mol. Biol.*, **1711**, 133–148.

Wadi,L. *et al.* (2016) Impact of outdated gene annotations on pathway enrichment analysis. *Nat. Methods*, **13**, 705–706.

Wang,J. *et al.* (2017) In-depth method assessments of differentially expressed protein detection for shotgun proteomics data with missing values. *Sci. Rep.-uk*, **7**, 3367.

Wertheim,G.B.W. *et al.* (2009) The Snf1-related kinase, Hunk, is essential for mammary tumor metastasis. *Proc. Natl. Acad. Sci.*, **106**, 15855–15860.

Wickham,H. (2009) *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer.

Zheng,Q. and Wang,X.-J. (2008) GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. *Nucleic Acids Res.*, **36**, W358–W363.